Hybrid Networks: Lessons Learned and Future Challenges Based on ESnet4 Experience

Inder Monga, Chin Guok, William E. Johnston, and Brian Tierney, Lawrence Berkeley National Laboratory/ESnet

ABSTRACT

ESnet, the Energy Sciences Network, has the mission of providing the network infrastructure to the U.S. Department of Energy's Office of Science programs and facilities, which depend on large collaborations and large-scale data sharing, enabling them to accomplish their science. ESnet4 — a hybrid IP and dynamic circuit network designed in 2006 and completed in 2008 — has managed to effectively satisfy the networking needs of the science community, easily handling dramatic growth in traffic requirements: around 80 percent growth year over year and 300 percent growth with the Large Hadron Collider (LHC) coming online. In this article, we examine the benefits and limitations of the current hybrid architecture based on actual production experience; discuss open research problems; and predict factors that will drive the evolution of hybrid networks, including advances in network technology, new computer architectures, and the onset of large-scale distributed computing.

INTRODUCTION

Across a number of scientific disciplines, users are demanding more network capabilities, including higher network end-to-end throughput, guaranteed bandwidth, and quality of service, in order to collaborate and share large amounts of data around the globe. In many fields, scientific discovery is now directly related to the amount of data that can be shared, accessed, and processed collaboratively.

Our understanding of the requirements of modern science for network services has emerged from detailed discussions with researchers[1] about how their analysis and simulation systems actually work: where the data originates, how many systems are involved in the analysis, how the systems and collaborators are distributed, how much data flows among these systems, how complex the work flow is, what the time sensitivitiesare , and so forth. Such applications are typically data-intensive and high-performance, frequently moving terabytes a day for months at a time; high-duty-cycle, operating most of the day for months at a time in order to meet requirements for data movement and analysis; and widely distributed, typically spread over continental or intercontinental distances. Considering the overall requirements, a set of generic but important goals were identified for any network and network services implementation [1]:

- Bandwidth: Adequate network capacity is required to ensure timely and high-performance movement of data produced by the facilities.
- Reliability: High reliability is required for large instruments and "systems of systems" (large distributed systems) that now depend on the wide area network (WAN) for inter-node communication.
- Connectivity: The network must have geographic reach — either directly or through peering arrangements with other networks — sufficient to connect users and collaborators and analysis systems to U.S. Department of Energy (DOE) Office of Science (SC) facilities.[2]
- Services: Guaranteed bandwidth, traffic isolation, end-to-end monitoring, and so on are required as network services, and these must be presented to the users in the context of service oriented architecture (SOA), the grid, and "systems of systems," which are the programming paradigms of modern science.

In addition, the nodes of the distributed systems must be able to get guarantees from the network that there is adequate network capacity over the entire lifetime of the task at hand. The systems must also be able to get performance and state information from the network to support graceful failure, auto-recovery, and adaptation to unexpected network conditions that are short of outright failure.

To address the above science needs, ESnet4 was designed as a hybrid network consisting of two core networks:

[1] *See the reports collected at http://www.es.net/requirements.html*

[2] *See "National Lab" and "User Facilities" tabs at http://www.sc.doe.gov/*

0163-6804/11/$25.00 © 2011 IEEE

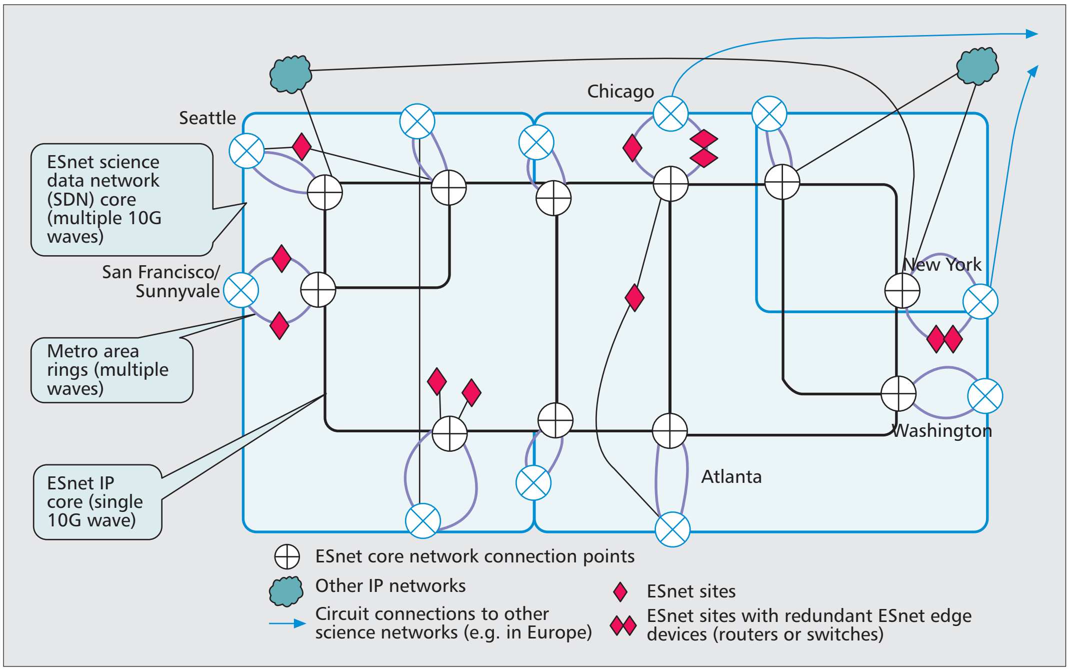

Seattle
Chicago
New York
Washington
Atlanta
San Francisco/ Sunnyvale
ESnet science data network (SDN) core (multiple 10G waves)
Metro area rings (multiple waves)
ESnet IP core (single 10G wave)
ESnet core network connection points
Other IP networks
Circuit connections to other science networks (e.g. in Europe)
ESnet sites
ESnet sites with redundant ESnet edge devices (routers or switches)

Figure 1. *ESnet4 architecture.*

ESnet is a national infrastructure with a richly interconnected topology built from multiple 10 Gb/s circuits connecting a collection of points of presence in major U.S. cities and at national R&E exchange points. The optical infrastructure covers most of the United States in six interconnected rings.

- An IP core that carries all the commodity IP traffic
- A circuit-oriented core (called the Science Data Network or SDN) primarily designed to carry large scientific data flows [1]

Both cores connect to peer research and education (R&E) networks in order to scale this architecture globally.

In this article, we present a brief overview of the ESnet4 hybrid network; discuss lessons learned, benefits realized, and unsolved challenges for the community; and briefly look at the future technologies and requirements that will drive the next-generation architecture of hybrid networking.

ESNET4: A BRIEF OVERVIEW

ESnet is a national infrastructure with a richly interconnected topology built from multiple 10 Gb/s circuits connecting a collection of points of presence (PoPs) in major U.S. cities and at national R&E exchange points. The optical infrastructure covers most of the United States in six interconnected rings. One 10 Gb/s footprint on the core network is dedicated to general IP traffic, and all other 10 Gb/s links are devoted to the SDN. At the current rate of increase, the SDN will use 40–60 Gb/s on most of the national network by 2011. Additionally, all of the DOE national laboratories are dually connected to the core, mostly by a collection of metro area optical rings in the San Francisco Bay, Chicago, and New York-Long Island area. Laboratories not in these metro areas are connected by loops off the core network (Fig. 1).

The On Demand Secure Circuits and Reservation System (OSCARS) virtual circuit service that serves the science applications is essentially a network management and control system for multiprotocol label switching;MPLS.[3] OSCARS supports routing constraints that are outside of the scope of the standard MPLS with Traffic Engineering (MPLS-TE) network configuration tools. In particular, due to the temporal nature of the circuits (reservation-based with a specified bandwidth, and start and end times), OSCARS manages a centralized temporal topology database that contains all of the link capacities and commitments over time. The link topology information is obtained using the Open Shortest Path First-Traffic Engineering (OSPF-TE) extension of the OSPF routing protocol used in the core network. Requests for new circuits are processed based on the link availability information in the topology database. Once a path is determined (assuming the request is consistent with available link capacities), it is set up link by link through the network using Resource Reservation Protocol with TE (RSVP-TE) to construct the MPLS label switched path (LSP) that defines the virtual circuit provided to the user.

At the data transport layer, which is the transport service offered to the user, the circuit can be established at layer 2 as a tagged Ethernet virtual local area network (VLAN) or at layer 3 as special routing applied to the IP address of the source (the science system).

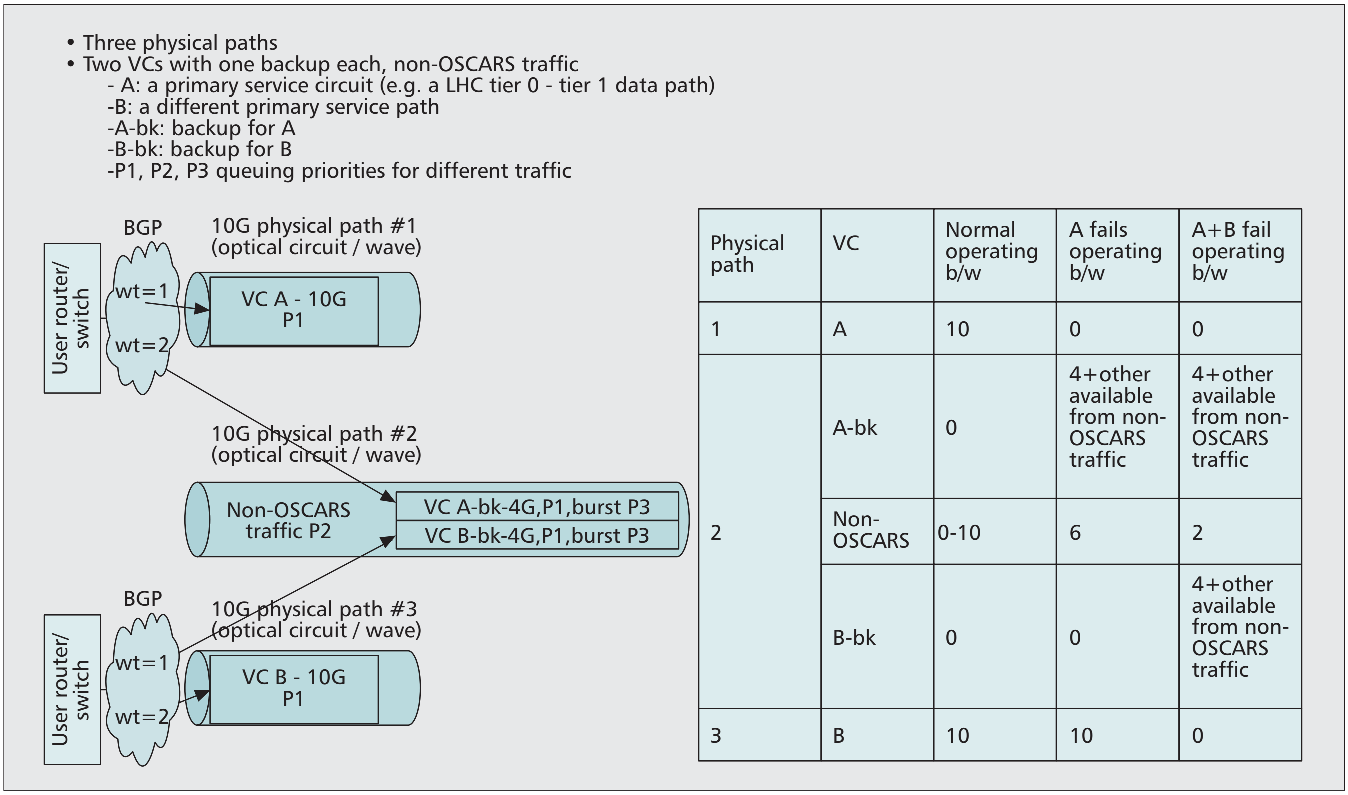
The bandwidth guarantees are provided:

- By assigning the circuit traffic an elevated queueing priority
- By doing admission control so that no link that carries OSCARS circuit traffic is ever oversubscribed
- By managing the traffic flowing into each virtual circuit

This ensures that the circuit traffic has priority over any other traffic on the link and that circuits do not interfere with each other.

The virtual circuits are rate-limited at the ingress but permitted to burst above the allocated bandwidth if idle capacity is available. This is accomplished without interfering with other cir-

[3] *MPLS is a highly scalable, protocol agnostic, data-carrying mechanism supported in many modern network routers.*



- Three physical paths
- Two VCs with one backup each, non-OSCARS traffic
    - A: a primary service circuit (e.g. a LHC tier 0 - tier 1 data path)
    -B: a different primary service path
    -A-bk: backup for A
    -B-bk: backup for B
    -P1, P2, P3 queuing priorities for different traffic

Physical path	VC	Normal operating b/w	A fails operating b/w	A+B fail operating b/w
1	A	10	0	0
2	A-bk	0	4+other available from non-OSCARS traffic	4+other available from non-OSCARS traffic
	Non-OSCARS	0-10	6	2
	B-bk	0	0	4+other available from non-OSCARS traffic
3	B	10	10	0

Figure 2. *Example of user-defined traffic engineering.*

cuits by marking the over-allocation bandwidth as low-priority traffic.

In order to provide high reliability, the user can request a second circuit that is diversely routed from the first circuit.

Together with user Border Gateway Protocol (BGP) sessions managing the input to the circuits, these semantics provide users with powerful user-level TE capability, allowing them to establish their own failover and repurposing and fair-sharing strategies among several different applications or users (Fig. 2).

The example of Fig. 2 is not speculative, but is routinely used by the two national laboratories (Brookhaven and Fermilab) that are Large Hadron Collider (LHC) Tier 1 data centers in the United States. The Tier 1 centers archive and redistribute a significant fraction of all of the data coming from the two largest LHC detectors (Atlas and CMS, the Tier 0 sources). This entails data flows from CERN to the Tier 1s that currently amount to a steady-state 10–30 Gb/s for about nine months a year. The data analysis is mostly done by Tier 2 centers at universities. The data transfers out of the Tier 1s to the Tier 2s are comparable, in aggregate, to the LHC (Tier 0) to Tier 1 transfers.

OSCARS is a production service in ESnet and currently manages 31 long-term circuits that serve four science disciplines: high-energy physics, climate research, computational astrophysics, and genomics. All LHC Tier 1 data paths within the United States utilize OSCARS circuits.

The final aspect of the virtual circuit service is that it is only useful if it provides end-to-end guaranteed throughput across multiple network domains, because essentially all science data flows originate in one domain (e.g., a national laboratory on ESnet) and terminate in another domain (e.g., a science group on a U.S. or European campus).

In order to provide this capability, a group of R&E networks has defined an Inter-Domain Control Protocol (IDCP) [2]. The IDCP has standardized the information and messages needed to set up end-to-end circuits across multiple domains, that is, for the exchange of topology information containing at least potential virtual circuit (VC) ingress and egress points, how to propagate the circuit setup request, and how data plane connections are facilitated across domain boundaries. Thus, OSCARS is an interdomain controller (IDC). It should be noted that while the IDCP does coordinate the specifics of the data plane technology that "stitches" two network domains together (e.g., Ethernet VLANs), it does not dictate how a virtual circuit (VC) is provisioned within a domain (e.g. Eo-MPLS). While OSCARS is fairly widely used in other network domains, there are several IDCs based on different approaches that have been demonstrated to interoperate within the United States and internationally [3]. Standardization of this approach is being undertaken within the Open Grid Forum (OGF) in the Network Services Interface (NSI) working group [4].

LESSONS LEARNED

OSCARS was first deployed in the ESnet3 backbone in 2005. In 2007 it began to manage all dynamic virtual circuits in the ESnet4 SDN core. Shortly after its deployment, several limitations quickly emerged that were not considered in the initial design of the service. These limitations mainly revolved around the disparity between the designed and actual use model of the service, service representation, and complex policy enforcement. Other operational requirements also surfaced that highlighted the need for transport efficiency, service resiliency, service transparency, monitoring, and brokering and co-scheduling.

Use Model Disparity

The initial design goal of the service was to have users schedule guaranteed bandwidth circuits only when needed, and return the resources otherwise. However, this was not the case due to three considerable issues:

- Timescale of minutes lead time needed to set up a circuit
- Lack of systems in place that could dynamically configure the user's local network
- Need for coordinating information such as Ethernet VLAN IDs and IP addresses between users that connected over an end-to-end circuit

Compared to typical control plane signaling mechanisms such as RSVP-TE, the provisioning workflow in an IDC is significantly more complex. This results in lead times of minutes for on-demand immediate provisioning, and is compounded for multidomain end-to-end circuits as IDCs in the path process the request sequentially.

OSCARS currently only manages the network resources in the WAN. However in order for an end-to-end circuit to function, "last mile" network elements (e.g., the campus LAN) must also be configured to carry the circuit through. Without an automated system, LAN administrators are hesitant to configure the devices manually to provide the last part of the circuit because this can be a tedious and error prone process.

Even if the last mile networks are configured dynamically, the end users that utilize the end-to-end circuit must coordinate among themselves to ensure connectivity. This may entail selecting a common Ethernet VLAN ID and/or IP addresses within the same subnet. This requires some knowledge of networking, which is not always the case for end users.

Due to these issues and how the circuits are used, almost all production circuits currently provisioned by OSCARS in ESnet have a lifespan of 12 months or more. This type of use has resulted in a different set of requirements needed to support such long-lived reservations, mainly how to support guaranteed bandwidth reservations and statistically multiplex them such that the network is efficiently used. To address this issue, it was determined that using the notion of a "floor" guarantee for circuits would be a good compromise. Essentially the user is allowed to send traffic that exceeds the reserved guaranteed bandwidth if excess capacity is available. In congestion scenarios, the user will be throttled back to the reserved guaranteed bandwidth. A practical example of this is the LHC Tier 1 to Tier 2 transfers, where up to 10 1 Gb/s circuits are reserved over common 10 Gb/s links. However, based on the actual usage by other users, each Tier 1 to Tier 2 transfer can potentially use the entire 10 Gb/s path at a time.

None of this precludes the use of short duration circuits (hours to days), and OSCARS can manage many of these.

Service Representation

The current reservation request message for a guaranteed bandwidth circuit requires the user to select things like source and destination points, bandwidth, and duration. These are reasonably straightforward parameters. However, the execution of the reservation is a circuit, which will have specific characteristics such as jitter and latency. The subtleties of these characteristics may not be obvious to non-network-savvy users. For example, someone who is interested in large data transfers is less concerned about latency and jitter than someone who wants to transport video from an instrument for remote control. The need to abstract the underlying network technology capabilities and present them in a service friendly manner is critical to the wide use and adoption of the system. Service abstraction and aggregation, and service selection can potentially be used to address this.

The need to abstract the underlying network technology capabilities and present them in a service friendly manner is critical to the wide use and adoption of the system. Service abstraction and aggregation, and service selection can potentially be used to address this.

Complex Policy Enforcement

In a dynamic circuit service, where scarce resources need to be managed, enforcement of complex service level agreement (SLA) policies can be nontrivial. In ESnet, it was determined that the following requirements were needed to support a user-driven dynamic circuit service:

- Enforcing service level expectations (SLEs) with trusted peers (intermediate network domains in a multidomain circuit) based on the identity of the peer and not the peer's customers. This removes the burden on ESnet to maintain peer customer information.
- Restricting which reservation endpoints can be selected based on the connectivity needs of a project. This prevents someone from illegitimately reserving bandwidth to arbitrary end sites.
- Creating functions that permit a LAN site administrator to view all site-related reservations and deny any unauthorized circuits to the site. This allows the LAN administrator to enforce site-specific acceptable user policies (AUPs).

All the issues described above have been addressed at some level in the OSCARS system. However, multiconstraint path finding algorithms can address these issues more effectively.

Transport Efficiency

The cost to move bits generally increases by a factor of two or more each time it moves up a network layer [5]. For example, the cost to move data at layer 3 is about 10 times the cost of moving it at layer 2, which in turn is a about twice the cost of moving it at layer 1. As such, an important practical requirement for dynamic circuit services is to intelligently determine the lowest network layer that can meet the service requirements of the user.

This functionality can be addressed using multiconstraint path finding algorithms.

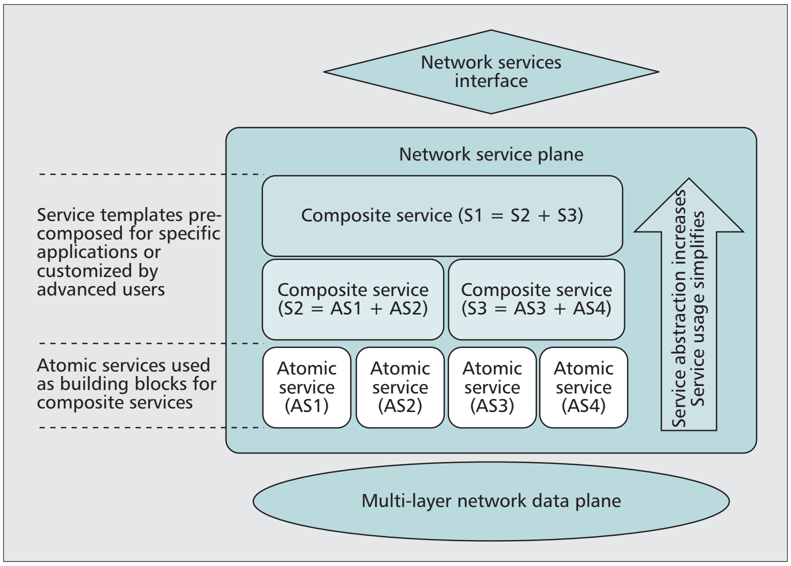

Network services interface
Network service plane
Service templates pre-composed for specific applications or customized by advanced users
Composite service (S1 = S2 + S3)
Composite service (S2 = AS1 + AS2)
Composite service (S3 = AS3 + AS4)
Atomic services used as building blocks for composite services
Atomic service (AS1)
Atomic service (AS2)
Atomic service (AS3)
Atomic service (AS4)
Service abstraction increases
Service usage simplifies
Multi-layer network data plane

Figure 3. *Composable services architecture.*

Service Resiliency

Service resiliency can easily be taken for granted in a best effort routing environment; however, for guaranteed bandwidth dynamic circuit services, this can be a hard problem. Within a single domain, protection and restoration schemes can be straightforward by basing it on the underlying transport technology. However, in a multidomain scenario, service resiliency may require the working and protect (backup) paths to traverse different domains for diversity. Coordinating a failover in this scenario requires someone with a global view of both paths, and this is typically the end user and not the individual network domain operators. Service control and management are needed to address this issue.

Service Transparency

Dynamically traffic engineering the network to adapt to load, reduce energy, and so on is an emerging trend. One solution to this is to utilize dynamic circuits within the network. However, to reduce the impact of traffic engineered network reconfigurations, the changes must be transparent to the user. Service control and management may potentially be enhanced to address this.

Monitoring

In a complex hybrid network that crosses multiple network domains, it is important to have a monitoring system that works across network domains and supports both active and passive monitoring. Our experience is that “soft failures” in multidomain paths are common and can be very time consuming to isolate. In soft failures packets still get delivered, but there is enough packet loss (not caused by congestion) to severely degrade TCP performance. Note that on a high-latency path even a 1 percent packet loss rate can lead to very poor TCP data transport performance.

To address this problem, we deployed the perfSONAR [6] suite of network monitoring tools, both within ESnet and between ESnet and several other R&E networks, including Internet2, NLR, and GEANT. perfSONAR is a web-services-based infrastructure for collecting and publishing network performance monitoring in a multidomain federated environment. The goal is to enable ubiquitous gathering and sharing of performance information to simplify management and facilitate cross-domain troubleshooting.

We are currently in the process of integrating perfSONAR with OSCARS to automatically monitor dynamic virtual circuits.

Brokering and Scheduling

Presenting the network as a schedulable resource, as opposed to just a transport mechanism, is extremely useful in promoting the model of infrastructure as a service (IaaS). To support this, it is necessary to understand how network resources can be brokered and co-scheduled. This is an ongoing research topic in several projects such as ORCA-BEN [7], GridARS [8], and GEYSERS [9].

Open Issues and Research Topics in Hybrid Networking

For hybrid networking to be successful, it must be end-to-end, simple to use, schedulable, and predictable. To accomplish this, there are several hard problems that still remain to be fully addressed. These include service abstraction and aggregation, service selection, multiconstraint path finding, and service control and management.

Service Abstraction and Aggregation

To make hybrid networking user friendly, the key is to abstract and hide the complexities of the underlying network technology from the user while at the same time providing sufficient “knobs” to allow sophisticated use of the service. It is necessary to map specific technology traits into terms that any user can understand. In a multidomain setting, certain end-to-end characteristics of a circuit may be easy to quantify. An example is latency, which is simply compounded over the domains in the path. Other traits, such as protection, may be very difficult to describe, because each domain along the end-to-end path may use different protection schemes based on the underlying transport technology. Solving this problem requires research on how networks can be characterized, normalized across different transport technologies, and mapped to user-friendly terms.

Service abstraction and aggregation are needed to address the requirement for user service representation.

Service Selection

Developing a good service model is essential for the adoption of hybrid networking. Having the ability to quantify services in the network at an atomic level and allowing users to build complex services using these atomic building blocks is a powerful concept. For example, a user moving a large file may just want a connection service to guarantee bandwidth between two points and a measurement service to determine its “goodput” (i.e., application-level throughput). Research on defining atomic services and a scalable composable service framework is needed. The framework should provide template services (e.g., bulk transport service, videoconferencing service) to novice users while supporting customizable composite services to expert users (Fig. 3).

Service selection is needed to address the requirement for user service representation.

Multiconstraint Path Finding

Multiconstraint path finding is essential to hybrid networking. It provides the necessary intelligence to determine how to efficiently reserve network resources while effectively meeting the user and network operator requirements. To facilitate complex multiconstraint path finding, the ARCHSTONE project [10] has implemented a flexible path computation engine (PCE) workflow model, whereby purpose-specific component PCEs are connected in a workflow graphs to incrementally prune network resources that do not meet the constraints of the user or network operator. This notion of having a network-holistic dedicated path computation process is consistent with the premise of the Internet Engineering Task Force (IETF) Path Computation Element Working Group. Consider the example PCE workflow in Fig. 4. In the PCE workflow graph, PCE1 may enforce administrative policy constraints and remove all links the user is not authorized to reserve. At this point, the pruned network resource graph is fed to two parallel sub-workflows: PCE2 → PCE3 → PCE4 and PCE5 → PCE6. In the first sub-workflow, PCE2 may enforce layer 3 bandwidth connectivity constraints and remove all links that are not in the potential path(s); PCE3 may enforce maximum latency constraints and remove links that exceed the user requested latency bounds; and PCE4 may enforce maximum jitter constraints and remove all network elements that exceed the user requested latency bounds. In the second sub-workflow, PCE5 may enforce layer 1 bandwidth connectivity constraints, while PCE6 may enforce latency constraints. (At layer 1 [e.g., optical], jitter characteristics are normally negligible; therefore, a jitter constraint PCE is unnecessary.) The resulting solutions from sub-workflow 1 and 2 are sent to the PCE aggregator to decide a solution to return to the user. (Note that any valid results from either sub-workflow 1 or 2 have met all the user requirements.)

The research focus in ARCHSTONE is to develop more efficient and purpose-driven PCEs and determine the optimal graph for the workflow.

Complex multiconstraint path finding is needed to address the requirements for complex policy enforcement and user service transparency.

Service Control and Management

Service control and management are needed to coordinate service functions over multiple domains, or dynamically adapt local network resources to changing requirements. In a multidomain scenario, end-to-end service functions need to be coordinated in order to make the service user friendly. For instance, if a user requests a working and protected circuit that requires some domain diversity (e.g., the working circuit traverses domains A→B→Z, and the protect circuit traverses A→C→Z). In the event of a failure in B there must be some coordination between A, C, and Z to switch the user transport from the working path onto the protect path. Similarly, in an end-to-end multilayer circuit where the data plane exchanges between domain boundaries at different layers (e.g., source→A at layer 3 [IP], A→B at layer 1 [e.g., SONET or optical], B→Z at layer 2 [Ethernet], and Z→sink at layer 3), the appropriate layering information (e.g., SONET time-division multiplex [TDM] time slots, Ethernet VLAN IDs) must be coordinated across the domains along the end-to-end path. Determining how this information is captured, preserved, and disseminated across the domains of interest is an open research topic.

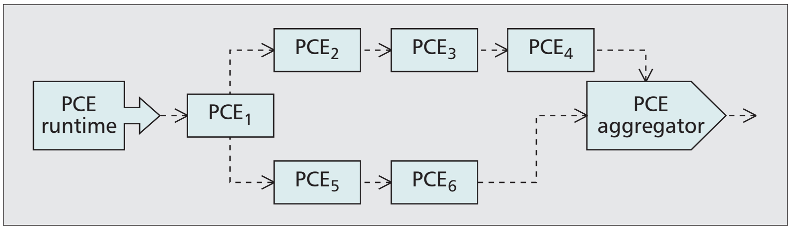


Figure 4. *PCE workflow example.*

Dynamically altering a network to adapt to changing requirements requires a fair amount of intelligent processing. Where multiconstraint path computation determines the "how," service control and management dictate the "when." Optimizing networks (which can be in several dimensions) in real time is a hard problem, and stands as a broad and open research topic.

Service control and management are needed to address the requirements for user service resiliency and service transparency.

Hybrid Networking: Next-Generation Requirements

The above research issues are being addressed within the context of the current applications and network technologies. This section discusses new applications and network innovations on the horizon that could have a major impact on hybrid networking and its evolution.

Higher Speed Transport: 100G, 400G, and Beyond

The current hybrid model of hard separation of bandwidth, with 10G links dedicated to be either commodity IP or guaranteed VCs, has worked well as traffic separation architecture in hybrid networks. With the advent of higher-speed physical interfaces, 100G and beyond [13], IP and VCs will most likely be carried over the same physical wavelength. Aggregation, de-aggregation, and dynamic movement of flows between layers could then be handled most efficiently within the fabric of a single device. Thus, the multilayer hybrid network, now implemented with different network elements handling pack-

The DOE has made a commitment to fund research to build an exascale computer by the end of this decade. Given the orders of magnitude difference in capabilities from existing computers, the HPC community is rethinking the basic architectural components to design a system to accomplish this goal.

ets or circuits, will likely collapse onto a single multilayer converged device. Current industry standards working in sub-rate bandwidth separation from coarse (e.g., optical paths) to granular (e.g., carrier Ethernet or MPLS) will likely be leveraged heavily to segregate the flows while providing service guarantees. Large flow identification and dynamic flow movement between the TDM/circuit and packet switching fabrics, as researched under the HNTES project [11], can be one of the service features that could automatically exploit the benefits of transport efficiencies at a particular layer while maintaining *service transparency*.

Cloud Architectures

Cloud architectures are being discussed and deployed in the commercial world, and are evolving from a special service offered by a few to a mix of private clouds, public clouds, special-purpose clouds (storage clouds, software as a service), and so on. "Cloud bursting," more specifically, dynamic allocation of resources in a public cloud to supplement a private cloud, is an important feature from which science users can benefit. Such inter-cloud interaction — movement of computation and data between various clouds — is predicted as one common mode of operation. In each of these scenarios, moving data to computation or computation to data in an expedient and dynamic manner is seen as a critical requirement. Thus, rapid provisioning and allocation of multidomain hybrid network resources to handle large ephemeral flows while ensuring user-friendly *service aggregation and abstraction* will be required to address the requirements. This application has the potential to change the use case within ESnet from only long-lived large flows to a mix with short-duration flows. This scenario may also place new requirements on the infrastructure/hardware and protocols to support sub-minute to subsecond provisioning of dynamic paths and network energy efficiency considerations [12].

Exascale Computing

The DOE has made a commitment to fund research to build an exascale computer by the end of this decade. Given the orders of magnitude difference in capabilities from existing computers, the high-performance computing (HPC) community is rethinking the basic architectural components to design a system to accomplish this goal. One of the key changes proposed is for internal communication between chips to transition from copper interfaces to optical. This low-level integration of optics on computer chips provides an interesting inflection point for hybrid networks. Given the high-throughput requirements of supercomputers — terabits per second between the supercomputing centers — the new architecture will have to deal with large bursts of information offered in the optical domain rather than as discrete packets today, and will demand seamless transport, potentially without buffering, from chip to chip over the wide area network. The challenge will be to create a *hybrid service control and management framework* to manage end-to-end transport. This would include end-to-end instantiation of features like sub-second protection and monitoring across multiple domains since applications may not be able to buffer information and recover from any connection failures at those high transport rates.

Conclusions

Building hybrid networks has served the needs of the scientific community well, and is now a proven architecture that could be adopted by commercial providers as they start dealing with dynamic large flows like HD videoconferencing, inter-cloud flows, data center traffic, video distribution, and caching. Lessons learned from the current deployment of ESnet's hybrid network have spawned open research problems that will drive automated allocation of network resources toward a path of maturity. Hybrid networks are well positioned to contribute to upcoming application areas like cloud computing and deal with radical innovations like exascale computing. Simplicity in service selection will be critical to expand the adoption of dynamic hybrid networking services by end users. Research and development activities thus need to continue to evolve the network service layer on top of the hybrid network.

Acknowledgment

This work was supported by the Director, Office of Science of the U.S. Department of Energy under Contract no. DE-AC02-05CH11231. Thanks to all current OSCARS open source contributors (http://code.google.com/p/oscars-idc/people/list), in particular, Evangelos Chaniotakis, Andy Lake, Eric Pouyoul, and Mary Thomson from ESnet, Jeff Boote and Aaron Brown at Internet2, and Tom Lehman and Xi Yang from ISI.

References

[1] W. Johnston *et al.*, "The Evolution of Research and Education Networks and their Essential Role in Modern Science," in *Trends in High Performance & Large Scale Computing*, L. Grandinetti and G. Joubert, Eds.; http://www.es.net/pub/esnet-doc/index.html.

[2] IDCP (2010), http://www.controlplane.net/.

[3] W. Johnston, E. Chaniotakis, and C. Guok, "ESnet and the OSCARS VIRTUAL CIRCUIT Service: Motivation, Design, Deployment and Evolution of a Guaranteed Bandwidth Network Service Supporting Large-Scale Science," http://www.es.net/pub/esnetdoc/index.html#oscars100510.

[4] Network Services Interface (NSI) Working Group, Open Grid Forum, http://ogf.org/gf/group_info/view.php?group=nsi-wg.

[5] C. de Laat, "The Power of Change," *OnVector 2010*, Feb. 8, 2010, p. 7; http://staff.science.uva.nl/~delaat/talks/cdl-2010-02-08.pdf.

[6] B. Tierney *et al.*, "perfSONAR: Instantiating a Global Network Measurement Framework," *SOSP Wksp. Real Overlays and Distrib. Sys.*, Oct. 2009.

[7] I. Baldine *et al.*, "The Missing Link: Putting the Network in Networked Cloud Computing," *Int'l. Conf. Virtual Computing Initiative*, North Carolina, Oct. 22–23, 2009.

[8] A. Takefusa *et al.*, "GridARS: An Advance Reservation-nased Grid Co-Allocation Framework for Distributed Computing and Network Resources," *Proc. 13th Int'l. Conf. Job Scheduling Strategies for Parallel Processing*, Seattle, WA, June 17, 2007.

[9] GEYSERS, 2010, http://www.geysers.eu/.

[10] ARCHSTONE, 2010, http://archstone.east.isi.edu/twiki/bin/view/ARCHSTONE/WebIndex.

[11] M. Veeraraghavan and Z. Yan, Hybrid Network Traffic Engineering Software; http://www.ece.virginia.edu/mv/research/DOE09/documents/deliverables/feb2010/mv-HNTES-final.pdf.

[12] J. Baliga *et al.*, "Green Cloud Computing: Balancing Energy in Processing, Storage and Transport," http://people.eng.unimelb.edu.au/rtucker/publications/files/Baliga_Ayre_Hinton_Tucker_JRLStrTrans.pdf.

[13] K. Roberts *et al.*, "100G and Beyond with Digital Coherent Signal Processing," *IEEE Commun. Mag.*, July 28, 2010, vol. 7, pp. 62–69.

Biographies

Inder Monga (imonga@es.net) is developing new ways to advance networking services for collaborative and distributed science by leading research and services within ESnet. He contributes to ongoing ESnet research projects, such as Advanced Resource Computation for Hybrid Service and Topology Nrtworks (ARCHSTONE) and ANI-Testbed as well as the OSCARS and Advanced Network Initiative (ANI) 100G projects. He also serves as co-chair of the NSIworking group in the Open Grid Forum. His research interests include network virtualization, network energy efficiency, grid/cloud computing, and sensor networking. He currently holds 10 patents and has over 15 years of industry and research experience in telecommunications and data networking at Wellfleet Communications, Bay Networks, and Nortel. He earned his undergraduate degree in electrical/electronics engineering from the Indian Institute of Technology, Kanpur, before coming to the United States to pursue his graduate studies in Boston University's Electrical and Electronics Communication Systems Department.

Chin Guok (chin@es.net) joined ESnet in 1997 as a network engineer, focusing primarily on network statistics. He was a core engineer in the testing and production deployment of MPLS and QoS (Scavenger Service) within ESnet. He is the technical lead of the ESnet On-Demand Secure Circuits and Advanced Reservation System (OSCARS) project, which enables end users to provision guaranteed bandwidth virtual circuits within ESnet. He also serves as a co-chair of the Open Grid Forum On-Demand Infrastructure Service Provisioning Working Group.

William E. Johnston (wej@es.net) is a senior scientist and advisor to ESnet, the network that serves the US Dept. of Energy Office of Science. He led ESnet between 2003 and 2008, during which time ESnet undertook a complete re-analysis of the requirements of the DOE's science programs that ESnet supports. As a result of this a new network architecture and an implementation approach were defined that would accommodate the massive data flows of science as typified by the movement of petabytes/year from the Large Hadron Collider (LHC). This new network was built in 2007 and 2008. Previously he ran the Lawrence Berkeley National Laboratory's Distributed Systems Department and worked on many projects related to the application of computing in science environments. He also co-founded the Grid Forum (now OGF) with Ian Foster and Charlie Catlett. He has worked in the field of computing for more than 40 years and has taught computer science at the undergraduate and graduate levels. He has a Master's degree in mathematics and physics from San Francisco State University. For more information see www.dsd.lbl.gov/ ~ wej.

Brian L. Tierney (bltierney@es.net) is a staff scientist and group leader of the ESnet Advanced Network Technologies Group at Lawrence Berkeley National Laboratory (LBNL). His research interests include high-performance networking and network protocols; distributed system performance monitoring and analysis; network tuning issues; and the application of distributed computing to problems in science and engineering. He has been the PI for several DOE research projects in network and Grid monitoring systems for data intensive distributed computing. He has an M.S. in computer science from San Francisco State University and a B.A. in physics from the University of Iowa. He has been at LBNL since 1990.

Simplicity in the service selection will be critical to expand the adoption of dynamic hybrid networking services by the end-users. Research and development activities, thus, need to continue to evolve the network service layer on top of the hybrid network.